ChinaXiv合作期刊

基于生成式对抗网络的画作的图像合成方法 *

赵宇欣，王 冠

(天津大学 数学学院，天津 300354)

摘 要：画作的图像合成旨在将两个不同来源的图像分别作为前景和背景融合在一起，这通常需要局部风格迁移。现有的算法过程繁琐且耗时，不能做到实时的图像合成。针对这一缺点，提出了基于生成式对抗网络(GAN)的前向生成模型(PainterGAN)。PainterGAN 的自注意力机制和 U-net 结构控制合成过程中前景的语义内容不变。同时，对抗学习保证逼真的风格迁移。在实验中，使用预训练模型作为 PainterGAN 的生成器，极大地节省了计算时间和成本。实验结果表明，比起已有的方法，PainterGAN 生成了质量相近甚至更好的图像，生成速度也提升了 400 倍，在解决局部风格迁移问题上是高质量、高效率的。

关键词：图像风格迁移；生成对抗网络；图像合成；自注意力机制

中图分类号：TP391.41 doi: 10.19734/j.issn.1001-3695.2020.03.0082

Painterly image composition based on generative adversarial net

Zhao Yuxin, Wang Guan

(*School of Mathematics*, *Tianjin University*, *Tianjin* 300354, *China*)

Abstract: Painterly image compositing aims to harmonize a foreground image inserted into a background painting, which is done by local style transfer. The chief drawback of the existing methods is the high computational cost, which makes real-time operation difficult. To overcome this drawback, this paper proposed a feed-forward model based on generative adversarial network (GAN) , called PainterGAN. PainterGAN introducesd a self-attention network and a U-net to control the semantic content in the generated image. Meanwhile, adversarial learning guaranteed a faithful transfer of style. PainterGAN also introduced a pre-trained network within the generator to extract features. This allowed PainterGAN to dramatically reduce training-time and storage. Experiments show that, compared to state-of-art methods, PainterGAN generated images hundreds of times faster with comparable or superior quality. Therefore, it is effective and efficient for local style transfer.

Key words: image style transfer; generative adversarial net; image compositing; self-attention

chinaXiv:202009.00062v1

0 引言

图像合成属于图像变换问题，目的是通过模型将一个简单的粘贴合成图像转变成一个融合为一体的图像。例如，将一个人像(前景)插入到一张照片(背景)中，图像合成期望将两者融合在一起，使得观察者以为这个人像本来就在照片中。因为前景和背景的光线，明亮，纹理等风格特征不同，简单的粘贴合成会造成不自然的视觉效果，可以被轻易判断为假的合成物。因此需要一个融合过程将背景的部分风格迁移到前景来，使它们的合成物在视觉上是统一协调的。针对照片的图像合成，不同的工作分别通过匹配前景和后景的统计特征，如直方图，均值方差[1]，协方差[2]等进行融合。针对画作的图像合成，Luan 等人[3]提出了基于 PatchMatch 和神经网络的局部风格迁移模型。本文也就这一问题提出新的思路。

与图像合成紧密相连的一个概念就是图像的风格迁移。随着深度学习[4,5]的进一步发展，Gatys 等人[6]提出神经风格迁移(neural style transfer-NST)，通过深度神经网络将油画的风格特征迁移到图像上，同时保留了图像本身的内容。考虑到 NST 的优化过程较为耗时，Johnson[7]和 Ulyanov[8]设计了快速前向生成模型，提高了图像生成的速度。在这之后，大量的工作[9,10]被提出，积极推动了这个领域的发展。目前这些工作都是针对全局的风格迁移问题，不适用于画作的图像合成，例如，粘贴一束花到梵高的油画作品星空中，一个理想的融合结果是，这束花具有和画中其他植物相似的风格，而不是集中夜空，山脉，人物所有风格于一体。

生成式对抗网络(GAN)[11]在 2014 年被提出，在很多图像问题上有令人印象深刻的表现。它由生成器和鉴别器组成，其中生成器试图生成与真实数据相似的图片，而鉴别器则尽力识别出这些生成的图片，直到它们达到纳什平衡。在这种状态下，生成器可以生成足够逼真的数据。cGAN[12]用卷积神经网络构造生成器和鉴别器，并用于解决图像方面的问题。IcGAN[13]将 GAN 和编码器结合起来在特征空间编辑图像属性，以控制图像的生成。CycleGAN[14]用双向映射的 GAN 模型来完成图像到图像的生成任务。Zhang 等人[16]将自注意力机制插入到 GAN 中，图像生成质量大幅提升。不同于对图像迭代优化的思路，这些模型极大地提升了图像的生成速度。但是生成的图像细节性不够，不同像素区域之间相关性不强。

本文基于 GAN 提出一个用于画作图像合成的全新模型 PainterGAN。通过对抗训练，损失函数驱动 PainerGAN 学习目标背景的明暗，色彩，纹理等风格特征，同时尽最大可能保留训练数据的语义内容不变。在训练完成后，将任意内容的前景图像输入到模型中，PainterAGN 都可以将其渲染成目标的背景风格，当渲染完成的前景图像贴入背景中时，能完全融入其中，令观看者无法判断该合成图像的真假，PainterGAN 以此完成从背景到前景的局部风格迁移。在这个过程中，一个关键点是原始的内容和逼真的风格之间的矛盾。当前景的内容被赋予较大的权重，迁移的风格通常与背景不一致，反之，当风格迁移更被重视，原本的内容会有一定程度的信息损失。基于单阶段的优化方案很难同时平衡二者，如 PatchMatch[15]。基于二阶段的优化方案通过粗糙—细节两

收稿日期：2020-03-14；**修回日期：**2020-05-03 **基金项目：**国家自然科学基金资助项目(91746107)

作者简介：赵宇欣(1995-)，女，山西晋中人，硕士研究生，主要研究方向为机器学习、深度学习、计算机视觉(zhaoyuxin_alice@tju.edu.cn)；王冠(1992-)，女，内蒙古呼伦贝尔人，博士研究生，主要研究方向为深度学习、数学物理反问题.

ChinaXiv合作期刊

chinaXiv:202009.00062v1

个阶段逐步优化生成图像，但是计算成本过高，如 DPH[3]。PainterGAN 在 GAN 的基础上作出改进，通过引入自注意力机制和 U-net 来控制前景的语义内容不变，同时对抗训练又保证风格逼真且与背景一致。在模型训练过程中 PainterGAN 用预训练的 VGG 替代生成器中的编码器，极大地节省了计算空间和时间。实验表明，本文模型在训练完成后可生成与现有模型质量相似甚至更好的图像，却将速度提高 400 倍。

1　本文方法

GAN 的基本思想是通过映射将特定数据分布变换为目标数据分布。训练过程中对抗损失函数驱动整个模型的参数优化，最终使之达到局部最优点。在画作的图像合成问题中，PainterGAN 的生成器将前景映射到背景图像的分布中，使之具有背景的风格特征。本节将对自注意力机制，PainterGAN 的网络结构和模型的损失函数进行详细描述。

1.1　自注意力机制的基本原理

自注意力机制在图像生成过程中通过建立不同像素区域的相关性，有助于促进物体的轮廓完整。在卷积计算中，比起整幅图像，单个卷积核通常提供很小的感受野，例如 3*3 或者 4*4。相应地，在卷积计算的前几层，图像的细颗粒度信息可以被捕获。随着层数的增加卷积核的感受野逐渐变大，模型能抓取图像中的语义内容，但是深层的特征映射丢失了很多信息，不同区域之间建立的联系很难有效传递到模型的浅层。因为卷积计算的这些局限，已有的风格迁移方法倾向于生成带有破碎边缘的物体。自注意力机制是一个可行的解决方法。

自注意力机制通常用于自然语言处理中的前后文语义理解。Zhang 等人[16]首次将其引入到 GAN 中用于图像分类。在其他的计算机视觉任务中，自注意力机制也被证明是有效的。从理论上说，它对人眼更容易注意到的图像区域反映更强烈，以此来增强物体的显著性。

自注意力网络被引入在 PainterGAN 的生成器下采样之后，上次采样之前。基本思想[16]可以总结为

a) 将编码器生成的特征映射输入到三个独立的卷积层，$f(x)=W_f x, g(x)=W_g x, h(x)=W_h x$ 中，假设输入为 $x\in R^{C*N}, N=h*w$，三个卷积层的系数矩阵分别为 $W_f\in R^{C*C}, W_g\in R^{C*C}, W_h\in R^{C*C}$。

b) $f(x_i), g(x_i)$ 用于计算 $h(x_i)$ 的权重：

$$\beta_{j,i}=\frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})}, s_{ij}=f(x_i)^T g(x_j) \tag{1}$$

其中 $\beta_{j,i}$ 用来衡量图像中第 i 块像素区域对生成第 j 块区域的重要性；

c) 输出是 $h(x_i)$ 的加权和 $o=(o_1,o_2,...,o_N)$：

$$o_j=\sum_{i=1}^{N}\beta_{j,i}h(x_i) \tag{2}$$

d) 考虑到一开始自注意力网络没有训练至局部最优点，参数 γ 用来调整输出：

$$y_i=\gamma o_i+x_i \tag{3}$$

通过以上步骤，自注意力网络逐步地发挥作用，来影响图像的生成。

1.2　PainterGAN 的网络结构

如图 1 所示，PainterGAN 主要包含两个部分，生成器和鉴别器。其中生成器由编码器和解码器构成，它们的网络结构对称，对输入的图像分别进行下采样和重构。为了节省计算空间和时间，PainterGAN 用训练好的 VGG-19 替代编码器。VGG 具有强大的特征提取功能，能同时抓取图像的像素级信息和语义内容。

下采样过程产生 32*32 的多通道特征映射，在其进入解码器之前，自注意力网络计算特征映射中不同区域的相关性。此外，U-net 串联编码器和解码器同一层级的特征。在编码器的指导下，解码器重构图像，生成器的结构如图 2 所示。

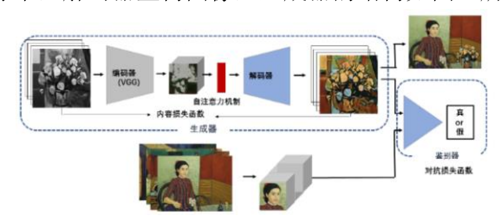


图 1　PainterGAN 的网络结构

Fig. 1　Overview of paintergan network structure

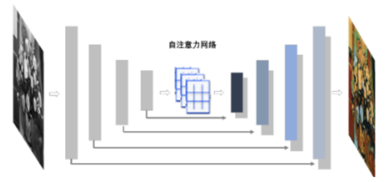


图 2　PainterGAN 中生成器的 U-net 结构

Fig. 2　The U-net structure of the generator in paintergan.

PainterGAN 的另一个重要组成部分是鉴别器。基于 GAN 的模型一般向鉴别器输入真实数据和生成数据。鉴别器经过一个简单的下采样过程，对输入数据给出真或(1)假(0)的判断。在本文的训练过程中，两种数据被分割为更小的像素块输入到模型的鉴别器中。这种处理减少了鉴别器的待训练参数，也使其有更高的灵活性，能够接收任何像素的图像作为输入。综上，鉴别器用于监督进行逼真的风格迁移，自注意力机制和 U-net 负责保存原本的语义内容，它们互相合作，保持风格迁移过程中内容和风格的平衡。

1.3　损失函数

1.3.1 对抗损失函数

正如前文中提到的，对抗损失函数驱动生成器和鉴别器达到平衡，二者的参数在训练过程中交替优化。生成器的损失函数为

$$L_{adv}(G)=E_{f_i\sim S_{data}(f)}[\log(1-D(G(f_i)))] \tag{4}$$

鉴别器的损失函数为

$$L_{adv}(D)=E_{g_i\sim S_{data}(g)}[\log(1-D(G(g_i)))]+E_{b_i\sim S_{data}(b)}[\log D(b_i)] \tag{5}$$

这里的 f_i, b_i, g_i 分别代表采样的前景图像，背景图像和生成图像。损失函数的值显示了生成图像在多大程度上拥有目标的风格。

1.3.2 内容损失函数

除去合理的风格，生成图像也应该保存其原本的语义内容。为了满足这个要求，DTN[17]发现当图像 x 经过生成器得到 $G(x)$，那么 $f(x)$ 和 $f(G(x))$ 是统一的，这里的 f 是指将图像映射到特征空间的函数，这种现象被称为“*f-constancy*”。它背后的逻辑是，外观改变的图像仍具有本身的高级语义内容特征。虽然这种方法是可行的，但是实验表明，“*f-constancy*”是一个过于严苛的限制，在一定程度上压制了风格的多样性。

本文采用了像素级的内容损失函数来衡量输入图像和生成图像在内容上的不同。为了得到更清晰的图像细节，采用 L1 范数计算：

$$L_{con}=E_{f_i\sim S_{data}(f)}[\|G(f_i)-f_i\|_1] \tag{6}$$

1.3.3 TV 正则项

为了鼓励图像的局部平滑，PainterGAN 采用了 TV 正则项：

$$L_{TV}=\sum_{i,j}((x_{i,j+1}-x_{i,j})+(x_{i+1,j}-x_{i,j})) \tag{7}$$

其中 $x_{i,j}$ 表示在(i, j)位置的像素值。综上，PainterGAN 的损失函数是

$$L(G,D)=\omega_1 L_{adv}+\omega_2 L_{con}+\omega_3 L_{TV} \tag{8}$$

其中 $\omega_1,\omega_2,\omega_3$ 分别代表对抗误差、内容误差和正则项在整个函

ChinaXiv合作期刊

数中的权重。

2　实验

2.1　实验平台信息

本文实验基于带有 NVIDIA GTX 1080 Ti GPU 处理器的 Ubuntu16.04 操作系统，通过 Python 语言和 Tensorflow 框架完成。预训练的 VGG-19 作为生成器中的编码器，生成的特征映射“*conv4_1*”作为自注意力网络的输入。U-net 连接对称的下采样和上采样卷积层。整个网络训练 200 个回合，每批 64 个数据。优化器为 Adam，初始学习率为 0.0002，动量为 0.5。

为了加速 PainterGAN 的生成器收敛到最优点，该网络被初始化为一个重构函数。只用内容损失函数训练 10 个回合，生成器即可生成与输入相近的图像。文献[18]也用同样的想法来加速模型优化。

2.2　实验数据与处理

本文训练数据包含两部分，灰度图像和呈现不同风格的画作。前者作为生成器的输入，后者和生成的图像作为鉴别器的输入。测试数据只包括灰度图像。

a) 前景。3482 张灰度图像来自电影《至爱梵高》，其中 3070 张用于训练模型，其余的用于测试。这些图像都经过裁剪，内容包括植物，建筑和人物等。

b) 背景。背景图像来自 4 部画作电影，其中《至爱梵高》共 2959 张，《父与女》共 2548 张，《回忆积木小屋》共 1570 张，《种树的牧羊人》共 4104 张。四组数据属于不同的画作流派，风格不同。所有的训练数据被裁剪至 256*256 大小，同时经过翻转和旋转来增强数据。

2.3　实验结果对比

实验对 NST[6]，Deep Analogy[19]，DPH[3]和 PainerGAN 进行对比，如图 3~7 所示。NST 通过对输入的白噪声不断优化进行全局风格迁移，实验结果显示这种方式并不适合画作的图像合成，合成物可以轻易被判断为假。例如，图 7 中的盘子混合了多种背景的颜色风格，使其与背景并不协调。此外，图 3 中的人物背景和图 5 中的火车都未能渲染合适的风格。

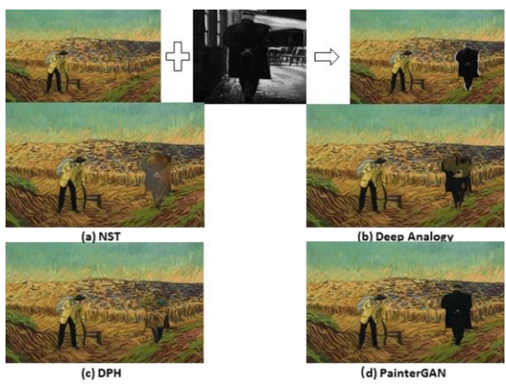


图 3　四种方法的实验结果对比图，背景来自《至爱梵高》

Fig. 3　Comparison of different approaches for compositing using background from Loving Vincent

DPH 和 Deep Analogy 的实验结果与本文 PainterGAN 的实验结果是具有可比性的，但是个别图片表现较差。例如，在图 4 中，DPH 将花瓶融入到了背景中，使前景的边缘线条难以分辨，这与整体的风格不一致。图 3 中的人物背景也有同样的问题。Deep Analogy 渲染前景的风格与背景是一致的，但是忽略前景的语义内容，因此只有在前景和后景内容相近的情况下表现较好。与它们相比，PainterGANGAN 不仅学习了视觉上足够逼真的风格，而且在合成过程中，考虑到了前景的语义内容，在此基础上进行合理的局部风格迁移。值得一提的是，本文的实验中前景统一被设置为 256*256，但是 PainterGAN 在测试过程中可以接收任意大小的图像作为输入。

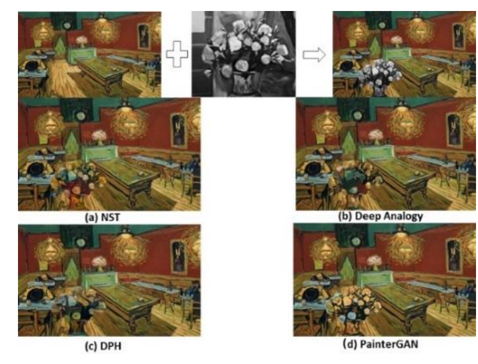


图 4　四种方法的实验结果对比图，背景来自《至爱梵高》

Fig. 4　Comparison of different approaches for compositing using background from Loving Vincent.

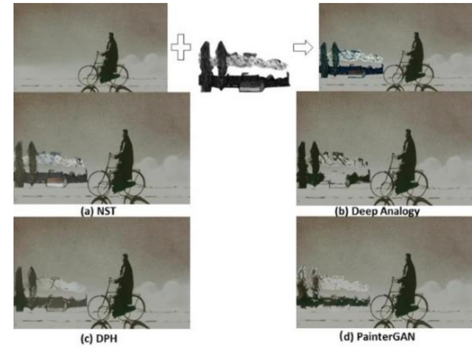


图 5　四种方法的实验结果对比图，背景来自《父与女》

Fig. 5　Comparison of different approaches for compositing using background from Father and Daughter.

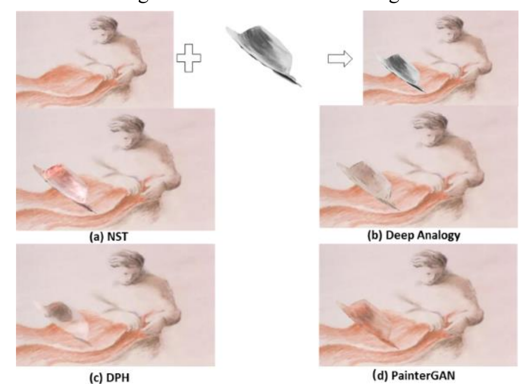


图 6　四种方法的实验结果对比图，背景来自《种树的牧羊人》

Fig. 6　Comparison of different approaches for compositing using background from The Man Who Planted Trees.



图 7　四种方法的实验结果对比图，背景来自《回忆积木小屋》

Fig. 7　Comparison of different approaches for compositing using background from The House of Small Cubes.

2.4　实验性能量化比较

在模型的训练时间方面，比较 PainterGAN 在未使用预训练的编码器和使用预训练的编码器两种情况下的计算时间和所需内存。结果显示，在达到相同的实验效果时，使用预训

chinaXiv:202009.00062v1

ChinaXiv合作期刊

chinaXiv:202009.00062v1

练的编码器减少了 33.82%的神经元，将模型的训练时间减少 46.49%。

在图像生成时间方面，四种风格迁移方法中，NST 是全局的风格迁移，Deep Analogy 对前景和后景有严格的要求。只有 DPH 和 PainterGAN 适用于对任意前景的局部风格迁移，因此这里对二者的图像合成时间进行比较，数值结果见表 1。

表 1　DPH 和 PainterGAN 的生成速度比较

Tab. 1　Comparison of generation time between DPH and paintergan

背景+前景	DPH	PainterGAN	提速倍数
至爱梵高+人物	7.50min	1s	**450x**
种树的牧羊人+帽子	6.97min	1s	**432x**
回忆积木小屋+盘子	7.42min	1s	**445x**
父与女+火车	7.57min	1s	**454x**
平均时间及倍数	7.37min	1s	**442x**

以上结果显示，PainterGAN 能实时地生成图像，比 DPH 快 400 倍。从这个角度看，PainterGAN 能有效学习图像的风格，并且能将任何前景高效地融入到该种风格的背景中。

2.5　损失函数的超参数调节

在 **1.3.3** 节中提到，PainterGAN 的损失函数为

$$L(G,D)=\omega_1 L_{adv}+\omega_2 L_{con}+\omega_3 L_{TV} \qquad (9)$$

其中超参数 $\omega_1,\omega_2,\omega_3$ 分别表示对抗误差项，内容误差项和正则项在驱动模型训练过程中的重要程度。通过多次实验和调节，$\omega_1,\omega_2,\omega_3$ 最终分别设为 1, 70, 50。这里选取 3 组不同超参数的损失函数和他们对应的测试结果进行对比，如图 8 所示。



图 8　三组不同超参数的测试结果对比，测试图像来自《至爱梵高》

Fig. 8　Comparison of three groups of loss function with different hyperparameters using image from Loving Vincent.

图 8 中(a)为输入的测试图像。(b)的损失函数中 $\omega_1:\omega_2:\omega_3$ 分别为 1:1:1， 可以看出图像的原始内容有部分丢失，如上图中的花瓣发生畸变，下图中花瓶的颈部图案丢失。于是在(c)中提高内容损失的权重，设 $\omega_1:\omega_2:\omega_3$ 为 1:50:1，但是在该组实验中，图中物体的边缘有不连续的情况，如下图中花朵的轮廓。(d)中相应提高 TV 项的权重，设 $\omega_1:\omega_2:\omega_3$ 为 1:70:50，抑制图像生成过程中的畸变，也保留了完整的语义内容，渲染效果最好。

3　结束语

PainterGAN 借助对抗训练，以图像到图像的前向生成方式重新考虑了图像合成中的局部风格迁移问题。它在 GAN 中引入自注意力机制和 U-net 来提高图像的生成质量，还进一步探索使用预训练的 VGG 作为生成器的编码器部分，在保持模型生成图像质量不变的情况下，节省了训练时间和内存。实验表明，比起已有的模型，PainterGAN 能完成与它们质量不相上下的风格迁移，甚至在某些情况下表现更好，同时极大地提高了图像的生成速度，实现了实时的图像合成。但是将本文的模型用在视频的局部风格迁移上仍有较大的问题，这也是未来一个值得研究的工作。

参考文献：

[1] Reinhard E, Ashikhmin M, Gooch B, *et al.* Color Transfer between Images [J]. IEEE Computer Graphics and Applications, 2001, 21 (5): 34-41.

[2] Li Yijun, Liu Mingyu, Li Xueting, *et al.* A closed-form solutionto photorealistic image stylization [C]// Proc of the European Conference on Computer Vision. Cham: Springer, 2018: 453-468.

[3] Luan Fujun, Paris S, Shechtman E, *et al.* Deep Photo Style Transfer [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscatawa, NJ: IEEE Press, 2017: 6997-7005.

[4] 刘建伟，刘媛，罗雄麟. 深度学习研究进展 [J]. 计算机应用研究, 2014, 31 (7): 1921-1930. (Liu Jianwei, Liu Yuan, Luo Xionglin. Research and development on deep learning [J]. Application Research of Computers, 2014, 31 (7): 1921-1930.)

[5] 毛勇华，桂小林，李前，等. 深度学习应用技术研究 [J]. 计算机应用研究, 2016, 33 (11): 3201-3205. (Mao Yonghua, Gui Xiaolin, Li Qian, *et al.* Study on application technology of deep learning [J]. Application Research of Computer, 2016, 33 (11): 3201-3205.)

[6] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style [J]. arXiv preprint arXiv: 1508. 06576, 2015.

[7] Johnson J, Alahi A, Li Fei Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 694-711.

[8] Ulyanov D, Lebedev V, Vedaldi A, *et al.* Texture Networks: Feed-forward Synthesis of Textures and Stylized Images [C]// Proc of the 33nd International Conference on Machine Learning. New York: ACM Press, 2016: 1349–1357.

[9] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// Proc of International Conference on Machine Learning. New York: ACM Press, 2015: 448–456.

[10] Li Chuan, Wand M. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscatawa, NJ: IEEE Press, 2016: 2479-2486.

[11] Goodfellow I J, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets [C]// International Conference on Neural Information Processing Systems. Boston: MIT Press, 2014: 2672-2680.

[12] Mirza M, Osindero S. Conditional Generative Adversarial Nets [J]. Computer Science, 2014, 5 (32): 2672-2680.

[13] Isola P, Zhu Junyan, Zhou Tinghui, *et al.* Image-to-Image Translation with Conditional Adversarial Networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscatawa, NJ: IEEE Press, 2016: 5967-5976.

[14] Zhu Junyan, Park T, Isola P, *et al.* Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscatawa, NJ: IEEE Press, 2017: 2242-2251.

[15] Connelly B, Eli S, Adam F, *et al.* PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing [J]. Acm Transactions on Graphics, 2009, 28 (3, article 24) .

[16] Zhang Han, Goodfellow I, Metaxas D, *et al.* Self-attention generative adversarial networks [J]. arXiv preprint arXiv: 1805. 08318, 2018.

[17] Taigman Y, Polyak A, Wolf L. Unsupervised cross-domain image generation [J]. arXiv preprint arXiv: 1611. 02200, 2016.

[18] Chen Yang, Lai Yu Kun, Liu Yong Jin. Cartoongan: Generative adversarial networks for photo cartoonization [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Piscatawa, NJ: IEEE Press, 2018: 9465-9474.

[19] Liao Jing, Yao Yuan, Yuan Lu, *et al.* Visual attribute transfer through deep image analogy [J]. Acm Transactions on Graphics, 2017, 36 (4): 1-15.